NPS55-87-010

NAVAL POSTGRADUATE SCHOOL

Monterey, California

TECHNICAL REPORT

PREDICTION OF POLYTOMOUS EVENTS: MODEL DESCRIPTION, ALGORITHM DEVELOPMENT AND METHODOLOGICAL ASPECTS, WITH AN APPLICATION

I. G. O'MUIRCHEARTAIGH
D. P. GAVER

SEPTEMBER 1987

Prepared for:
Naval Postgraduate School
Monterey, CA 93943

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral R. C. AUSTIN
Superintendent

K. T. MARSHALL
Acting Provost

This report was prepared by:

CURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NPS55-87-010	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Postgraduate School — 6b OFFICE SYMBOL (If applicable) Code 55	7a NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Monterey, CA 93943-5000	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING / SPONSORING ORGANIZATION — 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State and ZIP Code)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO / PROJECT NO / TASK NO / WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)
PREDICTION OF POLYTOMOUS EVENTS: MODEL DESCRIPTION, ALGORITHM DEVELOPMENT AND METHODOLOGICAL ASPECTS, WITH AN APPLICATION

12 PERSONAL AUTHOR(S)
O'MUIRCHEARTAIGH, I. G. and GAVER, D. P.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year Month Day)	15 PAGE COUNT
Technical	FROM ____ TO ____	1987 SEPT	19

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Generalized linear models, Logistics, Regression, Polytomou Prediction

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This article addresses the problem of predicting multi-(more than two) category events. It describes a new approach (based on generalized linear models) which is a natural extensi of the technique of logistic regression (a method widely used for the prediction of dichotomous events). As an example, the method is applied to the prediction of tropical storm imminence, and the performance of the method is compared with that of alternative well-established techniques.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Donald P. Gaver	22b TELEPHONE (Include Area Code) (408)646-2605 — 22c OFFICE SYMBOL Code 55 Gv

PREDICTION OF POLYTOMOUS EVENTS: MODEL DESCRIPTION, ALGORITHM DEVELOPMENT AND METHODOLOGICAL ASPECTS, WITH AN APPLICATION

I. O'MUIRCHEARTAIGH
D. P. GAVER

1. Introduction

The prediction of dichotomous events in meteorology (fog/no fog, precipitation/no precipitation) has been widely studied. Such predictions are also of interest in reliability and survival analysis, and in manpower planning. The analysis generally involves logistic regression or (equivalently) linear discriminant functions. Most standard statistical packages (e.g. BMDP, SAS) provide the facility for performing this analysis in some form. Also, the book by Cox (1970) can be consulted.

A natural extension of this problem (and one which has many potential applications in meteorology and elsewhere) is the situation in which the predictand is polytomous, i.e. has multiple categories. For example, it might be desired to predict visibility (good/marginal/bad) or precipitation (none/rain/snow). Two (methodologically) distinct cases can be envisaged, viz.

a. when the predictand involves ordered categories

b. when the categories are unordered.

Typically, the former case is the more common in meteorological applications, and this is the problem addressed in this report.

The particular application analyzed here involves 583 records of time to formation of tropical storms, and associated values of various meteorological variables. The time to storm formation is polytomized and recorded as

1: storm formed within 24 hours

2: storm formed between 24 and 48 hours

3: storm formed between 48 and 72 hours

4: storm did not form

The five meteorological variables recorded were:

X_1: unconditional probability of storm formation - a measure of the likelihood of storm formation in the particular disturbance location at the given time of year.

X_2: large scale vorticity (computed over 5 Latitude grid).

X_3: divergence (computed over 5 Latitude grid).

X_4: small scale vorticity (computed over 25 Latitude grid)

X_5: local generation of vorticity (product of X_3 and X_4).

Our objective was to determine how much predictive information is provided by the meteorological variables to facilitate prediction of imminence of tropical storms. Essentially the problem involves regression models where the dependent variable is ordinal. Much attention has been devoted to this general problem in the statistical literature of recent years (McCullagh [1980], McCullagh and Nelder [1983], Green [1984], Anderson [1984]). The central concept is that of the generalized linear model (McCullagh and Nelder [1983]).

In section 2 we describe briefly the concept of generalized models, and in more detail, the particular model (an extension of [dichotomous] logistic regression) utilized for our data. In section 3 we summarize the results of an ad hoc application of the model to our data, present the relevant parameter estimates, and evaluate the predictive performance of our model. A general discussion of our results is presented in section 4, together with suggestions for future work.

2. The Model.

2.1 General formulation.

Following Green [1984] and McCullagh and Nelder [1983], we consider a log likelihood L, a function of an n-vector, η, of predictors. We postulate, in our model, that the predictor η is functionally dependent on the p-vector β of parameters of interest. For our particular application, η, and β are specified in Section 2.2. The maximum likelihood estimation of β involves essentially solution of the equations

$$\frac{\delta L}{\delta \beta} = D^T u = 0 \qquad (1)$$

where $u_{nx1} = \frac{\delta L}{\delta \eta}$

and $D_{nxp} = \frac{\delta \eta}{\delta \beta}$

using the notation of Green [1984]. Again following Green, the standard Newton-Raphson method for the iterative solution of (1) involves evaluating u, D and the second derivatives of L for an initial value of β, and then solving the linear equations

$$\left(\frac{-\delta L}{\delta \beta \beta^T}\right) (\beta^* - \beta) = D^T u \qquad (2)$$

for an updated estimate β^*. Green shows that this is approximately equivalent to the solution of

$$(D^T A D)(\beta^* - \beta) = D^T u \tag{3}$$

where $A_{nxn} = E(\frac{\delta L}{\delta \eta} (\frac{\delta L}{\delta \eta})^T)$.

given an initial estimate of β (about which we have some further comments in Section 2.3). Equation (3) can be solved directly for β^*, or, equivalently, if a weighted least squares program is available, β^* results from regression $A^{-1}u + D\beta$ onto the columns of D using weight matrix A.

2.2 Specific Formulation.

In our application the data are in the form of N multinomial samples on the same set of k(=4) response categories (e.g. categories 1, 2, 3 and 4 indicating storm imminence as described in Section 1). The data may be arranged as a two-way table of counts y_{ij}, i=1...N; j=1, ...,4. The log-likelihood L is then given by

$$L = \sum_i \sum_j y_{ij} \log p_{ij} \tag{4}$$

where p_{ij} are the cell probabilities, and $\sum_{j=1}^{4} p_{ij} = 1$.

In the case where the categories 1,2,...,k are ordered, McCullagh and Nelder [1983] and Green [1984] both suggest the model

$$\eta_{ij} = \sum_{r=1}^{j} p_{ir} \equiv \psi\left(\frac{\theta_j - \Sigma x_{im}\beta_m}{\tau_i}\right) \tag{5}$$

$i=1, 2, \ldots N; \quad j=1,2,\ldots,k-1.$

where η_{ij} represent for fixed i, the cumulative cell probabilities, the matrix (x_{im}) represents covariate information and ψ is a given distribution function. Motivation for their model is provided by considering the response variable as an arbitrary grouping of an unobservable underlying variable on a continuous scale with "cutpoints" $\theta_1, \ldots \theta_{k-1}$. In some applications the θ_j's will be unknown and will need to be estimated; in others, such as the present one, they will be known, because the categorical variable (storm imminence coded 1,2,3,4) really is an arbitrary grouping of an underlying continuous variable based on known cutpoints (in this case that variable is time to storm formation with cut-points $\theta_1=24$, $\theta_2=48$ and $\theta_3=72$). For our application, we chose ψ to be the logistic distribution function, viz.,

$$\psi(x) = \frac{1}{1+e^{-x}} \tag{6}$$

This is the most widely used model in applications and has the advantage that a simple transformation can be used to (a) graphically check the suitability of the model and (b) provide initial values for the iterative estimation process.

Our model therefore is

$$\eta_{ij} = \frac{1}{1 + \exp - \dfrac{\theta_j - \sum_{m=1}^{M} x_{im}\beta_m}{\tau_i}} \tag{7}$$

or, in terms of previous notation

$$\eta = \eta(\beta) \tag{8}$$

where $\beta = (\beta_1, \ldots \beta_M, \tau_1, \ldots \tau_N)$. Unless we impose constraints, identifiability problems can arise. One common expedient (based on the concept of an underlying continuous variable used for the classification) is to allow an intercept and scale to write

$$\eta_{ij} = \psi\left(- \frac{[\theta_j - \gamma - \sum_{m=1}^{M} x_{im}\, \beta_m]}{\sigma}\right) \qquad i=1,\ldots N;\ j=1,\ldots,k-1. \tag{9}$$

where θ's are known.

A reformulation, more convenient for actual computation

$$\eta_{ij} = \psi(\beta_0 + \beta_1\theta_j + \sum_{m=1}^{M} x_{im}\beta_{m+2}) \tag{10}$$

2.3 Specific Methodology

We now apply the general methodology of Section 2.1 to the specific model described in equation (9). This involves two steps

(a) deriving explicit expressions for **A**, **D** and u described in that section, and

(b) Finding a suitable starting value β^* for the iterative reweighted least squares (IRLS) procedure.

We describe first the expressions for **D**, **A** and u for the special case of M=1 covariates with k=4 categories. The extension to other cases is straightforward. If we let

$$\eta_{nx1} = (\eta_{11}, \eta_{12}, \eta_{13}, \eta_{21}, \eta_{22}, \dots, \eta_{N3})$$

where n = Nx3, then $\mathbf{D} = \frac{\partial \eta}{\partial \beta}$ is a matrix given by

$$\mathbf{D}_{nx3} = \begin{pmatrix} F(\theta_1, x_1) & \theta_1 F(\theta_1, x_1) & x_1 F(\theta_1, x_1) \\ F(\theta_2, x_1) & \theta_2 F(\theta_2, x_1) & x_1 F(\theta_2, x_1) \\ F(\theta_3, x_1) & \theta_3 F(\theta_3, x_1) & x_1 F(\theta_3, x_1) \\ F(\theta_1, x_2) & \theta_1 F(\theta_1, x_2) & x_2 F(\theta_1, x_2) \\ \vdots & \vdots & \vdots \\ F(\theta_3, x_N) & \theta_3 F(\theta_3, x_N) & x_N F(\theta_3, x_N) \end{pmatrix} \tag{11}$$

where

$$F(\theta_i, x_j) = - \frac{\exp\{\beta_1 + \beta_2\theta_i + \beta_3 x_i\}}{\{1+\exp[\beta_1+\beta_2\theta_i+\beta_3 x_j]\}^2}$$

Similarly, $u_{nx1} = \frac{\partial L}{\partial \eta}$ is given by

$$
u = \begin{pmatrix}
\dfrac{y_{11}}{\eta_{11}} - \dfrac{y_{12}}{\eta_{12} - \eta_{11}} \\[2ex]
\dfrac{y_{12}}{\eta_{12}-\eta_{11}} - \dfrac{y_{13}}{\eta_{13} - \eta_{12}} \\[2ex]
\dfrac{y_{13}}{\eta_{13}-\eta_{12}} - \dfrac{y_{14}}{1 - \eta_{13}} \\[4ex]
\dfrac{y_{21}}{\eta_{21}} - \dfrac{y_{22}}{\eta_{22} - \eta_{21}} \\
\cdot \\
\cdot \\
\cdot \\
\dfrac{y_{N3}}{\eta_{N3}-\eta_{N2}} - \dfrac{y_{N4}}{1- \eta_{N3}}
\end{pmatrix} \tag{12}
$$

and $A_{nxn} = F\left(\dfrac{-\partial^2 L}{\partial\eta\eta^T}\right)$ is given by

$$
A = \left(\begin{array}{cccc}
\dfrac{-\eta_{12}}{\eta_{11}(\eta_{12}-\eta_{11})} & \dfrac{1}{\eta_{12}-\eta_{11})} & 0 & \\[2ex]
\dfrac{1}{\eta_{12}-\eta_{11}} & \dfrac{-(\eta_{13}-\eta_{11})}{(\eta_{12}-\eta_{11})(\eta_{13}-\eta_{12})} & \dfrac{1}{\eta_{13}-\eta_{12}} & 0 \\[2ex]
0 & \dfrac{1}{\eta_{13}-\eta_{12}} & \dfrac{-(1-\eta_{12})}{(\eta_{13}-\eta_{12})(1-\eta_{13})} & \\[1ex]
\hline
 & 0 & & \dfrac{-\eta_{22}}{\eta_{21}(\eta_{22}-\eta_{21})} \\
 & & & \text{etc.} \\
\hline
 & 0 & & 0
\end{array}\right) \tag{13}
$$

with tridiagonal 3x3 matrices similar to the one given above along the main diagonal, and zeroes elsewhere.

Each of **A**, **D** and **u** depends on the unknown parameters $\boldsymbol{\beta}$. Given an initial value for $\boldsymbol{\beta}$ we can evaluate **A**, **D** and **u** and commence the iterative process.

The initial values can be obtained by noting that if we apply a logit transformation to y_{ij} in equation (10) to give

$$\ell n \frac{\eta_{ij}}{1-\eta_{ij}} = \beta o + \beta_1\theta_j + \beta_2 x_i \tag{14}$$

then these logits are linear in the parameters $\underline{\beta}$. Initial estimates of $\underline{\beta}$ can be obtained by constructing empirical logits.

$$\ell_{ij} = \ln \left\{\frac{(\sum_{r=2}^{j} y_{ir}) + \frac{1}{2}}{n - (\sum_{r=1}^{j} y_{ir} + \frac{1}{2})}\right\} \tag{15}$$

and performing an unweighted least squares analysis for the model of equation (14) using these empirical logits as the dependent variable.

Two points should be noted in relation to the estimation of initial values for $\boldsymbol{\beta}$. Firstly, the IRLS procedure is quite sensitive to the choice of initial value, (see Green (1984)), and a poor choice can lead to non-convergence of the algorithm. Secondly, to obtain initial values by this procedure, it may be necessary to group the observations into categories based on values of the independent variables(s). If the data are not so grouped, (and our data consists essentially of multinomial samples of size 1), then all y_{ij} will be either 0 or 1 in our case this will lead to all the empirical logits having value either ln 3 or -ln 3.

Although, it may be necessary to group the data to obtain starting values for β, the maximum likelihood estimation of β may be carried out for either the grouped or the ungrouped data.

2.4 Computational procedure.

Since a stepwise program was not available, the model given by equation (10) was estimated separately for each covariate. Using the deviance (the likelihood ratio statistic against the saturated model) as a measure of goodness of fit, the best single explanatory variable was X_5. Having thus determined the optimal single variable for inclusion in the model, we then proceeded to establish which, if any, variable should next be included in the model. Due to the non-availability of a package for performing this analysis, the computation involved was cumbersome; the inclusion of an additional variable necessitated the re-programming of the computations leading to the matrices/vectors **A**, **D** and **u**. It was determined that X_2 was the next variable which should be included in the model. A third step of the stepwise procedure was also carried out, but no additional variable warranted inclusion in the model.

Accordingly, in evaluating the predictive performance of the model, and in comparing this performance with those of discriminant analysis and multiple regression, we used only the explanatory variables X_5 and X_2.

3. Evaluation of the predictive performance of the model.

3.1 Introduction

The model developed in this paper essentially produces, for given values of the covariates, estimates of η_{ij}, the cumulative category probabilities, and from these estimates of p_{ij}, the actual category probabilities. Hence, the model provides probability forecasts of the four storm imminence categories, for a given meteorological situation (as represented by the values of the meteorological variables). These probabilistic forecasts can be given directly as such, or may be converted, by methods described in Section 3.2, into categorical forecasts.

The problem of evaluating statistical forecasts of this type has been, and continues to be, a topic of major interest in meteorology. In Section 3.3, we consider two very simple methods of evaluating such forecasts; these two methods do not necessarily lead to the same conclusions in relation to the relative performance of the various forecasts.

3.2 Use of the Model for Forecasting

We consider two possibilities:

(a) Given the estimated probability forecast, a categorical forecast can be provided by forecasting the category of maximal probability. We denote this forecast by F1.

(b) The model described by equation (10) (or, more intuitively by equation (9)), suggests an underlying continuous variable, say Z, with the explanatory variable falling into categories 1,2,3 or 4 accordingly as $Z \leq \theta_1$, $\theta_1 < Z \leq \theta_2$, $\theta_2 < Z \leq \theta_3$, $Z > \theta_3$, respectively. Given values of the covariates, and estimates $\hat{\beta}$ of β, we can estimate Z by (in the notation of equation (10))

$$\hat{Z} = - \frac{\hat{\beta}_o + \sum_{m=1}^{M} X_m \hat{\beta}_{m+2}}{\hat{\beta}_1} \tag{17}$$

and then provide a categorical prediction that the storm imminence category is 1, 2, 3 or 4 according as

$$\hat{Z} \leq \theta_1 \quad \theta_1 < \hat{Z} \leq \theta_2, \quad \theta_2 < \hat{Z} \leq \theta_3, \quad \hat{Z} > \theta_3$$

This forecast is denoted by F2.

3.3 Evaluation of the forecasts

Since the two forecasts which we are considering here are categorical forecasts, one plausible criterion for evaluating these forecasts would be the number of correct forecasts achieved. Different forecasts can be readily compared using this measure. Since the categories being forecast are ordered, an incorrect forecast which is within one category of being correct is presumably preferable to one which "misses" by two or more categories. Hence an alternative measure of performance would be the number of forecasts which are within one category of being correct. We use both of the above measures in the paper to compare forecasts. As we will show, the different criteria can, in some cases, lead to a different ranking of forecasts.

In estimating the predictive performance of the model using the above measures, we omitted each data point in turn, estimated the model parameters from all the remaining data, and then used the forecast procedures F1 and F2 to predict the category of the omitted data point. For comparative purposes, we also used the standard techniques of discriminant analysis and multiple regression (using in the latter case as dependent variable the coded imminence of tropical storm variable, which takes values 1, 2, 3 and 4.

4. Discussion

The results of the cross-validation procedure, described in Section 3.3, for forecast methods F1 and F2, and for discriminant analysis and multiple regression, are given in Tables 1,2,3 and 4 respectively. An overall summary of the relative performance of the various methods is presented in Table 5. We would emphasise that any conclusions drawn here in relation to the efficacy of the various procedures are valid only in relation to the present application. Broader statements about the general performance of these methods would require extensive further analysis.

It is clear from Table 5 that no single technique is clearly superior. Using the criterion of maximising the number of correct forecasts, the generalized linear model applied in this paper, with forecasting strategy F2, is the best among those considered. However, if maximising the numbers of forecasts correct within one category is chosen as the comparative criterion, then multiple regression emerges as the optimal methodology.

From Table 5 it is clear that the performance of discriminant analysis is, in this application at least, somewhat inferior to that of the other techniques. A comparison of Tables 1 and 2 (which use the same model but different forecasting strategies based on the model) reveals some interesting facets about each of these strategies. Procedure F1 always forecasts either category 1 or category 4, and produces the highest overall percentage of correct forecasts. Procedure F2 is less extreme, and "smears" the category 1 forecasts over categories 1 and 2, and the category 4 forecasts over categories 3 and 4; this has the effect of reducing the percentage of correct forecasts but increasing the percentage of forecasts correct to within one category. Multiple regression "smears" the forecasts still further, with a resultant decrease in the percentage of correct forecasts and increase in the percentage correct to within one category. Multiple regression is, in fact, the method which produces the highest overall percentage of forecasts correct to within one category.

It is clear that the choice of strategy (forecasting procedure) among those considered and described here will be greatly influenced by the relative importance/seriousness of the various correct/incorrect forecasts. This strongly suggests that a decision theoretic approach might be considered, although, in practice, the specification of a loss function may be difficult and may unduly influence the choice of strategy.

Acknowledgements: I. G. O'Muircheartaigh and D. P. Gaver wish to express their gratitude to, respectively, the Naval Environmental Prediction Research Facility, Monterey, California and the Probability and Statistics Division of the Office of Naval Research, for their support in connection with this work.

Table 1
Forecast Procedure: F1

True State	Number of Forecasts				
	Forecast State				
	1	2	3	4	Total
1	27	0	0	34	61
2	10	0	0	32	42
3	4	0	0	20	24
4	11	0	0	445	456
total	52	0	0	531	583

Table 2
Forecast Procedure: F2

True State	Number of Forecasts				
	Forecast State				
	1	2	3	4	Total
1	18	9	7	27	61
2	8	2	5	27	42
3	1	3	3	17	24
4	4	7	17	428	456
total	31	21	32	499	583

Table 3
Forecast Procedure: Discriminant Analysis

	Number of Forecasts				
	Forecast State				
True State	1	2	3	4	Total
1	27	12	16	6	61
2	14	5	9	14	42
3	3	3	14	4	24
4	14	26	62	354	456
total	58	46	101	378	583

Table 4
Forecast Procedure: Multiple Regression

	Number of Forecasts				
	Forecast State				
True State	1	2	3	4	Total
1	4	21	31	5	61
2	0	11	21	10	42
3	0	2	18	4	24
4	0	11	110	335	456
total	4	405	180	354	583

Table 5

Forecast Procedure	% Correct Forecasts	% Forecasts Correct with one Category
F1	81	86
F2	77	88
Discriminant Analysis	69	86
Multiple Regression	63	90

5. References

Anderson, J. A. (1984), Regression and Ordered Categorical Variables, J. R. Statist. Soc., B,46, No. 1, pp. 1-30.

Cox, D. R. (1970), Analysis of Binary Data, London: Methuen.

Green, P. J., (1984), Iterated Rewighted Least Squares for Maximum Likelihood Estimation and Some Robust and Resistant Alternatives, J. R. Statist. Soc., B, 46, No. 2, pp. 149-192.

McCullagh P. (1980), Regression Models for Ordinal Data, J. R. Statist. Soc., B, 42, No. 2, pp. 109-142.

McCullagh P. and Nelder, J. A. (1983), Generalized Linear Models. London: Chapman and Hall.

DISTRIBUTION LIST

	NO. OF COPIES
Library (Code 0142) Naval Postgraduate School Monterey, CA 93943-5000	2
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Office of Research Administration (Code 012) Naval Postgraduate School Monterey, CA 93943-5000	1
Center for Naval Analyses 4401 Ford Ave. Alexandria, VA 22302-0268	1
Library (Code 55) Naval Postgraduate School Monterey, CA 93943-5000	1
Operations Research Center, Rm E40-164 Massachusetts Institute of Technology Attn: R. C. Larson and J. F. Shapiro Cambridge, MA 02139	1
Koh Peng Kong Ministry of Defense Blk A, Stockport Road SINGAPORE 0511	1